UTILITY PATENT APPLICATION TRANSMITTAL

(Only for new nonprovisional applications under 37 CFR 1.53(b))

Attorney Docket No.	M0656/7055
First Named Inventor or Application Identifier	
Venkataraman et al.	
Express Mail Label No.	EL056834527US
Date of Deposit	April 24, 2000

APPLICATION ELEMENTS

See MPEP chapter 600 concerning utility patent application contents

ADDRESS TO: **Box Patent Application**
Assistant Commissioner for Patents
Washington, DC 20231

1. ☐ Fee Transmittal Form
(Submit an original, and a duplicate for fee processing)

2. ☒ Specification [Total pages 35]
27 - pages specification
1 - pages abstract
7 - pages claims 53 - Total claims

3. ☒ Drawing(s) *(35 USC 113)* [Total sheets 5]
☒ Informal ☐ Formal [Total drawings **3**]

4. ☐ Oath or Declaration [Total pages]
a. ☐ Newly executed (original or copy)
b. ☐ Unsigned
c. ☐ Copy from a prior application (37 CFR 1.63(d))
(for continuation/divisional with Box 17 completed)
[Note Box 5 below]
i. ☐ DELETION OF INVENTOR(S)
Signed statement attached deleting inventor(s) named in the prior application, see 37 CFR 1.63(d)(2) and 1.33(b).

5. ☐ Incorporation by Reference
(usable if Box 4b is checked)
The entire disclosure of the prior application, from which a copy of the oath or declaration is supplied under Box 4b, is considered as being part of the disclosure of the accompanying application and is hereby incorporated by reference therein.

6. ☐ Microfiche Computer Program (Appendix)

7. ☐ Nucleotide and/or Amino Acid Sequence Submission (if applicable, all necessary)
a. ☐ Computer Readable Copy
b. ☐ Paper Copy (identical to computer copy)
c. ☐ Statement verifying identity of above copies

ACCOMPANYING APPLICATION PARTS

8. ☐ Assignment Papers/cover sheet & documents(s)

9. ☐ 37 CFR 3.73(b) Statement
(when there is an assignee)
☐ Power of Attorney

10. ☐ English Translation of Document *(if applicable)*

11. ☐ Information Disclosure Statement PTO-1449
☐ Copies of IDS Citations

12. ☐ Preliminary Amendment

13. ☒ Return Receipt Postcard (MPEP 503)
(Should be specifically itemized)

14. ☐ Small Entity Statement(s)
☐ Statement filed in prior application, Status still proper and desired

15. ☐ Certified Copy of Priority Document(s)
(if foreign priority is claimed)

16. Other:

17. If a CONTINUING APPLICATION, check appropriate box and supply the requisite information:

☐ Continuation ☐ Divisional ☐ Continuation-in-part (CIP) of prior application No.:

☐ Cancel in this application original claims of the prior application before calculating the filing fee.

☐ Amend the specification by inserting before the first line the sentence:

This application is a ☐ continuation ☐ divisional of application serial no. , filed , entitled , and now .

451667.1

18. CORRESPONDENCE ADDRESS					
Correspondence address below					
ATTORNEY'S NAME	Helen C. Lockhart, Reg. No. 39,248				
NAME	Wolf, Greenfield & Sacks, P.C.				
ADDRESS	600 Atlantic Avenue				
CITY	Boston	***STATE***	MA	***ZIP***	02210
COUNTRY	USA	***TELEPHONE***	(617) 720-3500	***FAX***	(617) 720-2441

19. SIGNATURE OF APPLICANT, ATTORNEY, OR AGENT REQUIRED	
NAME	Helen C. Lockhart, Reg. No. 39,248
SIGNATURE	Helen Lockhart
DATE	4/24/00

Express Mail Mailing Label Number EL056834527US
Date of Deposit April 24, 2000

SYSTEM AND METHOD FOR NOTATING POLYMERS

Related Application

This application claims priority under 35 U.S.C. §119 to US Provisional Patent Application Nos. 60/130,747, filed April 23, 1999, 60/130,792, filed April 23, 1999, 60/159,939, filed October 14, 1999, and 60/159,940, filed October 14, 1999, each of which are incorporated herein by reference in their entirety.

Background

Various notational systems have been used to encode classes of chemical units. In such systems, a unique code is assigned to each chemical unit in the class. For example, in a conventional notational system for encoding amino acids, a single letter of the alphabet is assigned to each known amino acid. A polymer of chemical units can be represented, using such a notational system, as a set of codes corresponding to the chemical units. Such notational systems have been used to encode polymers, such as proteins, in a computer-readable format. A polymer that has been represented in a computer-readable format according to such a notational system can be processed by a computer.

Conventional notational schemes for representing chemical units have represented the chemical units as characters (e.g., A, T, G, and C for nucleic acids), and have represented polymers of chemical units as sequences or sets of characters. Various operations may be performed on such a notational representation of a chemical unit or a polymer comprised of chemical units. For example, a user may search a database of chemical units for a query sequence of chemical units. The user typically provides a character-based notational representation of the sequence in the form of a sequence of characters, which is compared against the character-based notational representations of sequences of chemical units stored in the database. Character-based searching algorithms, however, are typically slow because such algorithms search by comparing individual characters in the query sequence against individual characters in the sequences of chemical units stored in the database. The speed of such algorithms is therefore related to the length of the query sequence, resulting in particularly poor performance for long query sequences.

Summary

In one aspect, the invention is directed to a notational system for representing polymers of chemical units. The notational system is referred to as Property encoded

nomenclature (PEN). According to one embodiment of the notational system, a polymer is assigned an identifier that includes information about properties of the polymer. For example, in one embodiment, properties of a disaccharide are each assigned a binary value, and an identifier for the disaccharide includes the binary values assigned to the properties of the disaccharide. In one embodiment, the identifier is capable of being expressed as a number, such as a single hexadecimal digit. The identifier may be stored in a computer readable medium, such as in a data unit (e.g., a record or a table entry) of a polymer database. Polymer identifiers may be used in a number of ways. For example, the identifiers may be used to determine whether properties of a query sequence of chemical units match properties of a polymer of chemical units. One application of such matching is to quickly search a polymer database for a particular polymer of interest or for a polymer or polymers having specified properties.

In one aspect, the invention is directed to a data structure, tangibly embodied in a computer-readable medium, representing a polymer of chemical units. In another aspect, the invention is directed to a computer-implemented method for generating such a data structure. The data structure may include an identifier that may include one or more fields for storing values corresponding to properties of the polymer. At least one field may be a non-character-based field. Each field may be capable of storing a binary value. The identifier may be a numerical identifier, such as a number that is representable as a single-digit hexadecimal number.

The polymer may be any of a variety of polymers. For example, (1) the polymer may be a polysaccharide and the chemical units may be saccharides; (2) the polymer may be a nucleic acid and the chemical units may be nucleotides; or (3) the polymer may be a polypeptide and the chemical units may be amino acids.

The properties may be properties of the chemical units in the polymer. For example, the properties may include charges of chemical units in the polymer, identities of chemical units in the polymer, confirmations of chemical units in the polymer, or identities of substituents of chemical units in the polymer. The properties may be properties of the polymer that are not properties of any individual chemical unit within the polymer. Example properties include a total charge of the polymer, a total number of sulfates of the polymer, a dye-binding of the polymer, a mass of the polymer, compositional ratios of substituents, compositional ratios of iduronic versus glucuronic, enzymatic sensitivity, degree of sulfation, charge, and chirality.

In another aspect, the invention is directed to a computer-implemented method for determining whether properties of a query sequence of chemical units match properties of a polymer of chemical units. The query sequence may be represented by a first data structure, tangibly embodied in a computer-readable medium, including an identifier that may include one or more bit fields for storing values corresponding to properties of the query sequence. The polymer may be represented by a second data structure, tangibly embodied in a computer-readable medium, including an identifier that may include one or more bit fields for storing values corresponding to properties of the polymer. The method may include acts of generating at least one mask based on the values stored in the one or more bit fields of the first data structure, performing at least one binary operation on the values stored in the one or more bit fields of the second data structure using the at least one mask to generate at least one result, and determining whether the properties of the query sequence match the properties of the polymer based on the at least one result. The chemical units may, for example, be any of the chemical units described above. Similarly, the properties may be any of the properties described above.

In one embodiment, the act of generating includes an act of generating the at least one mask as a sequence of bits that is equivalent to the values stored in the one or more bit fields of the first data structure. In another embodiment, the act of generating includes an act of generating the at least one mask as a sequential repetition of the values stored in the one or more bit fields of the first data structure.

In a further embodiment, the at least one mask includes a plurality of masks and the act of performing at least one binary operation includes acts of performing a logical AND operation on the values stored in the one or more bit fields of the second data structure using each of the plurality of masks to generate a plurality of intermediate results, and combining the plurality of intermediate results using at least one logical OR operation to generate the at least one result. In one embodiment, the act of determining includes an act of determining that the properties of the query sequence match the properties of the polymer when the at least one result has a non-zero value. In a further embodiment, the at least one binary operation includes at least one logical AND operation.

In another aspect, the invention is directed to a database, tangibly embodied in a computer-readable medium, for storing information descriptive of one or more polymers. The database may include one or more data units (e.g., records or table entries) corresponding to the one or more polymers, each of the data units may include an identifier

that may include one or more fields for storing values corresponding to properties of the polymer.

In another embodiment, the invention is directed to a data structure, tangibly embodied in a computer-readable medium, representing a chemical unit of a polymer. The data structure may comprise an identifier including one or more fields. Each field may be for storing a value corresponding to one or more properties of the chemical unit. At least one field may store a non-character-based value such as, for example, a binary or decimal value.

Other aspects of the invention include the various combinations of one or more of the foregoing aspects of the invention, as well as the combinations of one or more of the various embodiments thereof as found in the following detailed description or as may be derived therefrom. It should be understood that the foregoing aspects of the invention also have corresponding computer-implemented processes which are also aspects of the present invention. It should also be understood that other embodiments of the present invention may be derived by those of ordinary skill in the art both from the following detailed description of a particular embodiment of the invention.

Brief Description of the Drawings

FIG. 1 is a block diagram illustrating an example of a computer system for storing and manipulating polymer information.

FIG. 2A is a diagram illustrating an example of a record for storing information about a polymer and its constituent chemical units.

FIG. 2B is a diagram illustrating an example of a record for storing information about a polymer.

FIG. 2C is a diagram illustrating an example of a record for storing information about constituent chemical units of a polymer.

FIG. 3 is a flow chart illustrating an example of a method for determining whether properties of a first polymer of chemical units match properties of a second chemical unit.

Detailed Description

The present invention will be better understood in view of the following detailed description of a particular embodiment thereof, taken in conjunction with the attached drawings. All references cited herein are hereby expressly incorporated by reference.

FIG. 1 shows an example of a computer system 100 for storing and manipulating polymer information. The computer system 100 includes a polymer database 102 which

includes a plurality of records 104a-*n* storing information corresponding to a plurality of polymers. Each of the records 104a-*n* may store information about properties of the corresponding polymer, properties of the corresponding polymer's constituent chemical units, or both. The polymers for which information is stored in the polymer database 102 may be any kind of polymers. For example, the polymers may include polysaccharides, nucleic acids, or polypeptides.

A "polymer" as used herein is a compound having a linear and/or branched backbone of chemical units which are secured together by linkages. In some but not all cases the backbone of the polymer may be branched. The term "backbone" is given its usual meaning in the field of polymer chemistry. The polymers may be heterogeneous in backbone composition thereby containing any possible combination of polymer units linked together such as peptide- nucleic acids. In an embodiment, a polymer is homogeneous in backbone composition and is, for example, a nucleic acid, a polypeptide, a polysaccharide, a carbohydrate, a polyurethane, a polycarbonate, a polyurea, a polyethyleneimine, a polyarylene sulfide, a polysiloxane, a polyimide, a polyacetate, a polyamide, a polyester, or a polythioester. A "polysaccharide" is a biopolymer comprised of linked saccharide or sugar units. A "nucleic acid" as used herein is a biopolymer comprised of nucleotides, such as deoxyribose nucleic acid (DNA) or ribose nucleic acid (RNA). A polypeptide as used herein is a biopolymer comprised of linked amino acids.

As used herein with respect to linked units of a polymer, "linked" or "linkage" means two entities are bound to one another by any physicochemical means. Any linkage known to those of ordinary skill in the art, covalent or non-covalent, is embraced. Such linkages are well known to those of ordinary skill in the art. Natural linkages, which are those ordinarily found in nature connecting the chemical units of a particular polymer, are most common. Natural linkages include, for instance, amide, ester and thioester linkages. The chemical units of a polymer analyzed by the methods of the invention may be linked, however, by synthetic or modified linkages. Polymers where the units are linked by covalent bonds will be most common but also include hydrogen bonded, etc.

The polymer is made up of a plurality of chemical units. A "chemical unit" as used herein is a building block or monomer which can be linked directly or indirectly to other building blocks or monomers to form a polymer. The polymer preferably is a polymer of at least two different linked units. The particular type of unit will depend on the type of polymer. For instance DNA is a biopolymer comprised of a deoxyribose phosphate backbone composed of units of purines and pyrimidines such as adenine, cytosine, guanine,

thymine, 5-methylcytosine, 2-aminopurine, 2-amino-6-chloropurine, 2,6-diaminopurine, hypoxanthine, and other naturally and non-naturally occurring nucleobases, substituted and unsubstituted aromatic moieties. RNA is a biopolymer comprised of a ribose phosphate backbone composed of units of purines and pyrimidines such as those described for DNA but wherein uracil is substituted for thymidine. DNA units may be linked to the other units of the polymer by their 5' or 3' hydroxyl group thereby forming an ester linkage. RNA units may be linked to the other units of the polymer by their 5', 3' or 2' hydroxyl group thereby forming an ester linkage. Alternatively, DNA or RNA units having a terminal 5', 3' or 2' amino group may be linked to the other units of the polymer by the amino group thereby forming an amide linkage.

Whenever a nucleic acid is represented by a sequence of letters it will be understood that the nucleotides are in 5'→3' order from left to right and that "A" denotes adenosine, "C" denotes cytidine, "G" denotes guanosine, "T" denotes thymidine, and "U" denotes uracil unless otherwise noted.

The chemical units of a polypeptide are amino acids, including the 20 naturally occurring amino acids as well as modified amino acids. Amino acids may exist as amides or free acids and are linked to the other units in the backbone of the polymers through their a-amino group thereby forming an amide linkage to the polymer.

A polysaccharide is a polymer composed of monosaccharides linked to one another. In many polysaccharides the basic building block of the polysaccharide is actually a disaccharide unit which can be repeating or non-repeating. Thus, a unit when used with respect to a polysaccharide refers to a basic building block of a polysaccharide and can include a monomeric building block (monosaccharide) or a dimeric building block (disaccharide).

A "plurality of chemical units" is at least two units linked to one another.

The polymers may be native or naturally-occurring polymers which occur in nature or non-naturally occurring polymers which do not exist in nature. The polymers typically include at least a portion of a naturally occurring polymer. The polymers can be isolated or synthesized *de novo*. For example, the polymers can be isolated from natural sources e.g. purified, as by cleavage and gel separation or may be synthesized e.g.,(i) amplified *in vitro* by, for example, polymerase chain reaction (PCR); (ii) synthesized by, for example, chemical synthesis; (iii) recombinantly produced by cloning, etc.

Fig. 2A illustrates an example of the format of a data unit 200 in the polymer database 102 (i.e., one of the data units 104a-*n*). As shown in FIG. 2A, the data unit 200

may include a polymer identifier (ID) 202 that identifies the polymer corresponding to the data unit 200. The polymer ID 202 is described in more detail below with respect to FIG. 2B. The data unit 200 also may include one or more chemical unit identifiers (IDs) 204a-*n* corresponding to chemical units that are constituents of the polymer corresponding to the data unit 200. The chemical unit IDs 204a-*n* are described in more detail below with respect to FIG. 2C. The format of the data unit 200 shown in FIG. 2A is merely an example of a format that may be used to represent polymers in the polymer database 102. Polymers may be represented in the polymer database in other ways. For example, the data unit 200 may include only the polymer ID 202 or may only include one or more of the chemical unit IDs 204a-*n*.

FIG. 2B illustrates an example of the polymer ID 202. The polymer ID 202 may include one or more fields 202a-*n* for storing information about properties of the polymer corresponding to the data unit 200 (FIG. 2A). Similarly, FIG. 2C illustrates an example of the chemical unit 204a. The chemical unit ID 204a may include one ore more fields 206a-*m* for storing information about properties of the chemical unit corresponding to the chemical unit ID 204a. Although the following description refers to the fields 206a-*m* of the chemical unit ID 204a, such description is equally applicable to the fields 202a-*n* of the polymer ID 202a (and the fields of the chemical unit IDs 204b-*n*).

The fields 206a-*m* of the chemical unit ID 204a may store any kind of value that is capable of being stored in a computer readable medium, such as, for example, a binary value, a hexadecimal value, an integral decimal value, or a floating point value.

Each field 206a-*m* may store information about any property of the corresponding chemical unit. A "property" as used herein is a characteristic (e.g., structural characteristic) of the polymer that provides information (e.g., structural information) about the polymer. When the term property is used with respect to any polymer except a polysaccharide the property provides information other than the identity of a unit of the polymer or the polymer itself. A compilation of several properties of a polymer may provide sufficient information to identify a chemical unit or even the entire polymer but the property of the polymer itself does not encompass the chemical basis of the chemical unit or polymer.

When the term property is used with respect to polysaccharides, to define a polysaccharide property, it has the same meaning as described above except that due to the complexity of the polysaccharide, a property may identify a type of monomeric building block of the polysaccharide. Chemical units of polysaccharides are much more complex than chemical units of other polymers, such as nucleic acids and polypeptides. The

polysaccharide unit has more variables in addition to its basic chemical structure than other chemical units. For example, the polysaccharide may be acetylated or sulfated at several sites on the chemical unit, or it may be charged or uncharged. Thus, one property of a polysaccharide may be the identity of one or more basic building blocks of the polysaccharides.

A basic building block alone, however, may not provide information about the charge and the nature of substituents of the saccharide or disaccharide. For example, a building block of uronic acid may be iduronic or glucuronic acid. Each of these building blocks may have additional substituents that add complexity to the structure of the chemical unit. A single property, however, may not identify such additional substitutes charges, etc., in addition to identifying a complete building block of a polysaccharide. This information, however, may be assembled from several properties. Thus, a property of a polymer as used herein does not encompass an amino acid or nucleotide but does encompass a saccharide or disaccharide building block of a polysaccharide.

A type of property that provides information about a polymer may depend on a type of polymer being analyzed. For instance, if the polymer is a polysaccharide, properties such as charge, molecular weight, nature and degree of sulfation or acetylation, and type of saccharide may provide information about the polymer. Properties may include, but are not limited to, charge, chirality, nature of substituents, quantity of substituents, molecular weight, molecular length, compositional ratios of substituents or units, type of basic building block of a polysaccharide, hydrophobicity, enzymatic sensitivity, hydrophilicity, secondary structure and conformation (i.e., position of helicies), spatial distribution of substituents, ratio of one set of modifications to another set of modifications (i.e., relative amounts of 2-O sulfation to N-sulfation or ratio of iduronic acid to glucuronic acid), and binding sites for proteins. Other properties may be identified by those of ordinary skill in the art. A substituent, as used herein is an atom or group of atoms that substitute a unit, but are not themselves the units.

A property of a polymer may be identified by any means known in the art. The procedure used to identify a property may depend on a type of property. Molecular weight, for instance, may be determined by several methods including mass spectrometry. The use of mass spectrometry for determining the molecular weight of polymers is well known in the art. Mass Spectrometry has been used as a powerful tool to characterize polymers because of its accuracy (±1Dalton) in reporting the masses of fragments generated (e.g., by enzymatic cleavage), and also because only pM sample concentrations are required. For

example, matrix-assisted laser desorption ionization mass spectrometry (MALDI-MS) has been described for identifying the molecular weight of polysaccharide fragments in publications such as Rhomberg, A. J. et al, *PNAS, USA*, v. 95, p. 4176-4181 (1998); Rhomberg, A. J. et al, *PNAS, USA*, v. 95, p. 12232-12237 (1998); and Ernst, S. et. al., *PNAS, USA*, v. 95, p. 4182-4187 (1998), each of which is hereby incorporated by reference. Other types of mass spectrometry known in the art, such as, electron spray-MS, fast atom bombardment mass spectrometry (FAB-MS) and collision-activated dissociation mass spectrometry (CAD) can also be used to identify the molecular weight of the polymer or polymer fragments.

The mass spectrometry data may be a valuable tool to ascertain information about the polymer fragment sizes after the polymer has undergone degradation with enzymes or chemicals. After a molecular weight of a polymer is identified, it may be compared to molecular weights of other known polymers. Because masses obtained from the mass spectrometry data are accurate to one Dalton (1D), a size of one or more polymer fragments obtained by enzymatic digestion may be precisely determined, and a number of substituents (i.e., sulfates and acetate groups present) may be determined. One technique for comparing molecular weights is to generate a mass line and compare the molecular weight of the unknown polymer to the mass line to determine a subpopulation of polymers which have the same molecular weight. A "mass line" as used herein is an information database, preferably in the form of a graph or chart which stores information for each possible type of polymer having a unique sequence based on the molecular weight of the polymer. Thus, a mass line may describe a number of polymers having a particular molecular weight. A two-unit nucleic acid molecule (i.e., a nucleic acid having two chemical units) has 16 (4 units ^{2}) possible polymers at a molecular weight corresponding to two nucleotides. A two-unit polysaccharide (i.e., disaccharide) has 32 possible polymers at a molecular weight corresponding to two saccharides. Thus, a mass line may be generated by uniquely assigning a particular mass to a particular length of a given fragment (all possible di, tetra, hexa, octa, up to a hexadecasaccharide), and tabulating the results (An Example is shown in Figure 4).

Table 1 below shows an example of a computed set of values for a polysaccharide. From Table 1, a number of chemical units of a polymer may be determined from the minimum difference in mass between a fragment of length n+1 and a fragment of length n. For example, if the repeat is a disaccharide unit, a fragment of length n has 2n

monosaccharide units. For example, n=1 may correspond to a length of a disaccharide and n=2 may correspond to a length of a tetrasaccharide, etc.

Fragment Length n	Minimum difference in mass between n+1 and n (D)
1	101.13
2	13.03
3	13.03
4	9.01
5	9.01
6	4.99
7	4.99
8	0.97
9	0.97

TABLE 1

Because mass spectrometry data indicates the mass of a fragment to 1D accuracy, a length may be assigned uniquely to fragment by looking up a mass on the mass line. Further, it may be determined from the mass line that, within a fragment of particular length higher than a disaccharide, there is a minimum of 4.02D different in masses indicating that two acetate groups (84.08D) replaced a sulfate group (80.06D). Therefore, a number of sulfates and acetates of a polymer fragment may be determined from the mass from the mass spectrometry data and, such number may be assigned to the polymer fragment.

In addition to molecular weight, other properties may be determined using methods known in the art. The compositional ratios of substituents or chemical units (quantity and type of total substituents or chemical units) may be determined using methodology known in the art, such as capillary electrophoresis. A polymer may be subjected to an experimental constraint such as enzymatic or chemical degradation to separate each of the chemical units of the polymers. These units then may be separated using capillary electrophoresis to determine the quantity and type of substituents or chemical units present in the polymer. Additionally, a number of substituents or chemical units can be determined using calculations based on the molecular weight of the polymer.

In the method of capillary gel-electrophoresis, reaction samples may be analyzed by small-diameter, gel-filled capillaries. The small diameter of the capillaries (50 μm) allows for efficient dissipation of heat generated during electrophoresis. Thus, high field strengths can be used without excessive Joule heating (400 V/m), lowering the separation time to about 20 minutes per reaction run, therefor increasing resolution over conventional gel electrophoresis. Additionally, many capillaries may be analyzed in parallel, allowing amplification of generated polymer information.

In addition to being useful for identifying a property, compositional analysis also may be used to determine a presence and composition of an impurity as well as a main property of the polymer. Such determinations may be accomplished if the impurity does not contain an identical composition as the polymer. To determine whether an impurity is present may involve accurately integrating an area under each peak that appears in the electrophoretogram and normalizing the peaks to the smallest of the major peaks. The sum of the normalized peaks should be equal to one or close to being equal to one. If it is not, then one or more impurities are present. Impurities even may be detected in unknown samples if at least one of the disaccharide units of the impurity differs from any disaccharide unit of the unknown.

If an impurity is present, one or more aspects of a composition of the components may be determined using capillary electrophoresis. Because all known disaccharide units may be baseline-separated by the capillary electrophoresis method described above and because migration times typically are determined using electrophoresis (i.e., as opposed to electroosmotic flow) and are reproducible, reliable assignment to a polymer fragment of the various saccharide units may be achieved. Consequently, both a composition of the major peak and a composition of a minor contaminant may be assigned to a polymer fragment. The composition for both the major and minor components of a solution may be assigned as described below.

One example of such assignment of compositions involves determining the composition of the major AT-III binding HLGAG decasaccharide (+ DDD4-7) and its minor contaminant (+ D5D4-7) present in solution in a 9:1 ratio. Complete digestion of this 9:1 mixture with a heparinases yields 4 peaks: three representative of the major decasaccharide (viz., D, 4, and -7) which are also present in the contaminant and one peak, 5, that is present only in the contaminant. In other words, the area of each peak for D, 4, and -7 represents an additive combination of a contribution from the major decasaccharide and

the contribution from the contaminant, whereas the peak for 5 represents only the contaminant.

To assign the composition of the contaminant and the major component, the area under the 5 peak may be used as a starting point. This area represents an area under the peak for one disaccharide unit of the contaminant. Subtracting this area from the total area of 4 and -7 and subtracted twice this area from an area under D yields a 1:1:3 ratio of 4:-7:D. Such a ratio confirms the composition of the major component and indicates that the composition of the impurity is two Ds, one 4, one -7 and one 5.

Methods of identifying other types of properties may be easily identifiable to those of skill in the art and may depend on the type of property and the type of polymer. For example, hydrophobicity may be determined using reverse-phase high-pressure liquid chromatography (RP-HPLC). Enzymatic sensitivity may be identified by exposing the polymer to an enzyme and determining a number of fragments present after such exposure. The chirality may be determined using circular dichroism. Protein binding sites may be determined by mass spectrometry, isothermal calorimetry and NMR. Enzymatic modification (not degradation) may be determined in a similar manner as enzymatic degradation, i.e., by exposing a substrate to the enzyme and using MALDI-MS to determine if the substrate is modified. For example, a sulfotransferase may transfer a sulfate group to an HS chain having a concomitant increase in 80Da. Conformation may be determined by modeling and nuclear magnetic resonance (NMR). The relative amounts of sulfation may be determined by compositional analysis or approximately determined by raman spectroscopy.

FIG. 2D illustrates an example of the chemical unit ID 204a. The chemical unit ID 204a contains one or more fields 212a-e for storing information about properties of a heparin-like glycosaminoglycan (HLGAG). HLGAGs are complex polysaccharide molecules made up of disaccharide repeat units comprising hexoseamine and glucuronic/iduronic acid that are linked by α/β 1-4 glycosidic linkages. These defining units may be modified by: sulfation at the N, 3-O and 6-O position of the hexoseamine, 2-O sulfation of the uronic acid, and C5 epimerization that converts the glucuronic acid to iduronic acid. The disaccharide unit of HLGAG may be represented as:

$$(\alpha\ 1\rightarrow 4)\ I/G_{2OX}\ (\alpha/\beta\ 1\rightarrow 4)\ H_{3OX,NY}{}^{6OX}\ (\alpha\ 1\rightarrow 4),$$

where X may be sulfated ($-SO_3H$) or unsulfated (-H), and Y may be sulfated ($-SO_3H$) or acetylated ($-COCH_3$) or, in rare cases, neither sulfated nor acetylated.

The fields 212a-e may store any kinds of values, such as, for example single-bit values, single-digit hexadecimal values, or decimal values. In one embodiment, the chemical unit ID 204a includes each of the following fields: (1) a field 212a for storing a value indicating whether the polymer contains an iduronic or a glucuronic acid (I/G); (2) a field 212b for storing a value indicating whether the 2X position of the iduronic or glucuronic acid is sulfated or unsulfated; (3) a field 212c for storing a value indicating whether the hexoseamine is sulfated or unsulfated; (4) a field 212d indicating whether the 3X position of the hexoseamine is sulfated or unsulfated; and (5) a field 212e indicating whether the NX position of the hexoseamine is sulfated or acetylated. Optionally, each of the fields 212a-e may be represented as a single bit.

Table 2 illustrates an example of a data structure having a plurality of entries, where each entry represents an HLGAG encoded in accordance with Fig. 2D. Bit values for each of the fields 212a-e may be assigned in any known manner. For example, with respect to field 212a (I/G), a value of one may indicate Iduronic and a value of zero may indicate Glucuronic, or vice versa.

I/G	2X	6X	3X	NX	ALPH CODE	DISACC	MASS (ΔU)
0	0	0	0	0	0	I-H_{NAc}	379.33
0	0	0	0	1	1	I-H_{NS}	417.35
0	0	0	1	0	2	I-$H_{NAc,3S}$	459.39
0	0	0	1	1	3	I-$H_{NS,3S}$	497.41
0	0	1	0	0	4	I-$H_{NAc,6S}$	459.39
0	0	1	0	1	5	I-$H_{NS,6S}$	497.41
0	0	1	1	0	6	I-$H_{NAc,3S,6S}$	539.45
0	0	1	1	1	7	I-$H_{NS,3S,6S}$	577.47
0	1	0	0	0	8	I_{2S}-H_{NAc}	459.39
0	1	0	0	1	9	I_{2S}-H_{NS}	497.41
0	1	0	1	0	A	I_{2S}-$H_{NAc,3S}$	539.45
0	1	0	1	1	B	I_{2S}-$H_{NS,3S}$	577.47
0	1	1	0	0	C	I_{2S}-$H_{NAc,6S}$	539.45
0	1	1	0	1	D	I_{2S}-$H_{NS,6S}$	577.47

I/G	2X	6X	3X	NX	ALPH CODE	DISACC	MASS (ΔU)
0	1	1	1	0	E	I_{2S}-$H_{NAc,3S,6S}$	619.51
0	1	1	1	1	F	I_{2S}-$H_{NS,3S,6S}$	657.53
1	0	0	0	0	-0	G-H_{NAc}	379.33
1	0	0	0	1	-1	G-H_{NS}	417.35
1	0	0	1	0	-2	G-$H_{NAc,3S}$	459.39
1	0	0	1	1	-3	G-$H_{NS,3S}$	497.41
1	0	1	0	0	-4	G-$H_{NAc,6S}$	459.39
1	0	1	0	1	-5	G-$H_{NS,6S}$	497.41
1	0	1	1	0	-6	G-$H_{NAc,3S,6S}$	539.45
1	0	1	1	1	-7	G-$H_{NS,3S,6S}$	577.47
1	1	0	0	0	-8	G_{2S}-H_{NAc}	459.39
1	1	0	0	1	-9	G_{2S}-H_{NS}	497.41
1	1	0	1	0	-A	G_{2S}-$H_{NAc,3S}$	539.45
1	1	0	1	1	-B	G_{2S}-$H_{NS,3S}$	577.47
1	1	1	0	0		G_{2S}-$H_{NAc,6S}$	
1	1	1	0	1	-D	G_{2S}-$H_{NS,6S}$	577.47
1	1	1	1	0	-E	G_{2S}-$H_{NAc,3S,6S}$	619.51
1	1	1	1	1	-F	G_{2S}-$H_{NS,3S,6S}$	657.53

TABLE 2

Representing a HLGAG using a bit field may have a number of advantages. Because a property of an HLGAG may have one of two possible states, a binary bit is ideally-suited for storing information representing an HLGAG property. Bit fields may be used to store such information in a computer readable medium (e.g., a computer memory or storage device), for example, by packing multiple bits (representing multiple fields) into a single byte or sequence of bytes. Furthermore, bit fields may be stored and manipulated quickly and efficiently by digital computer processors, which typically store information

using bits and which typically can quickly perform operations (e.g., shift, AND, OR) on bits. For example, as described in more detail below, a plurality of properties each stored as a bit field can be searched more quickly than searches conducted using typical character-based searching methods.

Further, using bit fields to represent properties of HLGAGs permits a user to more easily incorporate additional properties (e.g., 4-O sulfation vs. unsulfation) into a chemical unit ID 204a by adding extra bits to represent the additional properties.

In one embodiment, the four fields 212b-e (each of which may store a single-bit value) may be represented as a single hexadecimal (base 16) number where each of the fields 212a-e represents one bit of the hexadecimal number. Using hexadecimal numbers to represent disaccharide units is convenient both for representation and processing because hexadecimal digits are a common form of representation used by conventional computers.

Optionally, the five fields 212a-e of the record 210 may be represented as signed hexadecimal digit, in which the fields 212b-212e collectively encode a single-digit hexadecimal number as described above and the I/G field is used as a sign bit. In such a signed representation, the hexadecimal numbers 0-F may be used to code chemical units containing iduronic acid and the hexadecimal numbers -0 to -F may be used to code units containing glucuronic acid. The chemical unit ID 204a may, however, be encoded using other forms of representations, such as by using a twos-complement representation.

The fields 212a-e of the chemical unit ID 204a may be arranged in any order. For example, a gray code system may be used to code HLGAGs. In a gray code numbering scheme, each successive value differs from the previous value only in a single bit position. For example, in the case of HLGAGs, the values representing HLGAGs may be arranged so that any two neighboring values differ in the value of only one property. An example of a gray code system used to code HLGAGs is shown in Table 3.

I/G 16	2X 8	6X 4	3X 2	NX 1	Numeric Value	DISACC	MASS (ΔU)
0	0	0	0	0	0	I-H_{NAc}	379.33
0	0	0	0	1	1	I-H_{NS}	417.35
0	0	0	1	1	3	I-$H_{NS,3S}$	497.41
0	0	0	1	0	2	I-$H_{NAc,3S}$	459.39
0	0	1	1	0	6	I-$H_{NAc,3S,6S}$	539.45

I/G 16	2X 8	6X 4	3X 2	NX 1	Numeric Value	DISACC	MASS (ΔU)
0	0	1	1	1	7	I-$H_{NS,3S,6S}$	577.47
0	0	1	0	1	5	I-$H_{NS,6S}$	497.41
0	0	1	0	0	4	I-$H_{NAc,6S}$	459.39
0	1	1	0	0	12	I_{2S}-$H_{NAc,6S}$	539.45
0	1	1	0	1	13	I_{2S}-$H_{NS,6S}$	577.47
0	1	1	1	1	15	I_{2S}-$H_{NS,3S,6S}$	657.53
0	1	1	1	0	14	I_{2S}-$H_{NAc,3S,6S}$	619.51
0	1	0	1	0	10	I_{2S}-$H_{NAc,3S}$	539.45
0	1	0	1	1	11	I_{2S}-$H_{NS,3S}$	577.47
0	1	0	0	1	9	I_{2S}-H_{NS}	497.41
0	1	0	0	0	8	I_{2S}-H_{NAc}	459.39
1	1	0	0	0	24	G_{2S}-H_{NAc}	459.39
1	1	0	0	1	25	G_{2S}-H_{NS}	497.41
1	1	0	1	1	27	G_{2S}-$H_{NS,3S}$	577.41
1	1	0	1	0	26	G_{2S}-$H_{NAc,3S}$	539.45
1	1	1	1	0	30	G_{2S}-$H_{NAc,3S,6S}$	619.51
1	1	1	1	1	31	G_{2S}-$H_{NS,3S,6S}$	657.53
1	1	1	0	1	29	G_{2S}-$H_{NS,6S}$	577.47
1	1	1	0	0	28	G_{2S}-$H_{NAc,6S}$	539.45
1	0	1	0	0	20	G-$H_{NAc,6S}$	459.39
1	0	1	0	1	21	G-$H_{NS,6S}$	497.41
1	0	1	1	1	23	G-$H_{NS,3S,6S}$	577.47
1	0	1	1	0	22	G-$H_{NAc,3S,6S}$	539.45
1	0	0	1	0	18	G-$H_{NAc,3S}$	459.39
1	0	0	1	1	19	G-$H_{NS,3S}$	497.41
1	0	0	0	1	17	G-H_{NS}	417.35
1	0	0	0	0	16	G-H_{NAc}	379.33

TABLE 3

Table 3 illustrates that use of a gray coding scheme arranges the disaccharide building blocks such that neighboring table entries differ from each other only in the value

of a single property. One advantage of using gray codes to encode HLGAGs is that a biosynthesis of HLGAG fragments may follow a specific sequence of modifications starting from the basic building block G-H_{HNac}.

In Table 3, bit weights of 8, 4, 2, and 1 are used to calculate the numerical equivalent of a hexadecimal number with the most significant bit (I/G) being used as a sign bit. For example, the hexadecimal code A (01010 binary) is equal to 8*1 + 4*0 + 2*1 + 1*0 = 10.

In another embodiment, the weights of each of the fields 212a-e may be changed thereby implementing an alternative weighting system. For example, bit fields 212a-e may have weights of 16, 8, 4, -2, and -1, respectively, as shown in Table 4.

I/G 16	2X 8	NX 4	3X -2	6X -1	Value	DISACC	MASS (ΔU)
0	0	0	0	0	0	I-H_{NAc}	379.33
0	0	0	0	1	-1	I-$H_{NAc,6S}$	459.39
0	0	0	1	0	-2	I-$H_{NAc,3S}$	459.39
0	0	0	1	1	-3	I-$H_{NAc,3S,6S}$	539.45
0	0	1	0	0	4	I-H_{NS}	417.35
0	0	1	0	1	3	I-$H_{NS,6S}$	497.41
0	0	1	1	0	2	I-$H_{NS,3S}$	497.41
0	0	1	1	1	1	I-$H_{NS,3S,6S}$	577.47
0	1	0	0	0	8	I_{2S}-H_{NAc}	459.39
0	1	0	0	1	7	I_{2S}-$H_{NAc,6S}$	539.45
0	1	0	1	0	6	I_{2S}-$H_{NAc,3S}$	539.45
0	1	0	1	1	5	I_{2S}-$H_{NAc,3S,6S}$	619.51
0	1	1	0	0	12	I_{2S}-H_{NS}	497.41
0	1	1	0	1	11	I_{2S}-$H_{NS,6S}$	577.47
0	1	1	1	0	10	I_{2S}-$H_{NS,3S}$	577.47
0	1	1	1	1	9	I_{2S}-$H_{NS,3S,6S}$	657.53
1	0	0	0	0	16	G-H_{NAc}	379.33
1	0	0	0	1	15	G-$H_{NAc,6S}$	459.39
1	0	0	1	0	14	G-$H_{NAc,3S}$	459.39
1	0	0	1	1	13	G-$H_{NAc,3S,6S}$	539.45

I/G 16	2X 8	NX 4	3X -2	6X -1	Value	DISACC	MASS (ΔU)
1	0	1	0	0	20	G-H_{NS}	417.35
1	0	1	0	1	19	G-$H_{NS,6S}$	497.41
1	0	1	1	0	18	G-$H_{NS,3S}$	497.41
1	0	1	1	1	17	G-$H_{NS,3S,6S}$	577.47
1	1	0	0	0	24	G_{2S}-H_{NAc}	459.39
1	1	0	0	1	23	G_{2S}-$H_{NAc,6S}$	539.45
1	1	0	1	0	22	G_{2S}-$H_{NAc,3S}$	539.45
1	1	0	1	1	21	G_{2S}-$H_{NAc,3S,6S}$	619.51
1	1	1	0	0	28	G_{2S}-H_{NS}	497.41
1	1	1	0	1	27	G_{2S}-$H_{NS,6S}$	577.47
1	1	1	1	0	26	G_{2S}-$H_{NS,3S}$	577.47
1	1	1	1	1	25	G_{2S}-$H_{NS,3S,6S}$	657.53

TABLE 4

Modifying the weights of the bits may be used to score the disaccharide units. For example, a database of sequences may be created and the different disaccharide units may be scored based on their relative abundance in the sequences present in the database. Some units, for example, I-$H_{NAc,3S}{}^{6S}$, which rarely occur in naturally-occurring HLGAGs, may receive a low score based on a scheme in which the bits are weighted in the manner shown in Table 4.

Optionally, the sulfation and acetylation positions may be arranged in an shown in Table 2: I/G, 2X, 6X, 3X, NX. These positions may, however, be arranged differently, resulting in a same set of codes representing different disaccharide units. Table 5, for example, shows an arrangement in which the positions are arranged as I/G, 2X, NX, 3X, 6X.

I/G	2X	NX	3X	6X	ALPH CODE	DISACC	MASS (ΔU)
0	0	0	0	0	0	I-H_{NAc}	379.33
0	0	0	0	1	1	I-$H_{NAc,6S}$	459.39
0	0	0	1	0	2	I-$H_{NAc,3S}$	459.39

I/G	2X	NX	3X	6X	ALPH CODE	DISACC	MASS (ΔU)
0	0	0	1	1	3	I-$H_{NAc,3S,6S}$	539.45
0	0	1	0	0	4	I-H_{NS}	417.35
0	0	1	0	1	5	I-$H_{NS,6S}$	497.41
0	0	1	1	0	6	I-$H_{NS,3S}$	497.41
0	0	1	1	1	7	I-$H_{NS,3S,6S}$	577.47
0	1	0	0	0	8	I_{2S}-H_{NAc}	459.39
0	1	0	0	1	9	I_{2S}-$H_{NAc,6S}$	539.45
0	1	0	1	0	A	I_{2S}-$H_{NAc,3S}$	539.45
0	1	0	1	1	B	I_{2S}-$H_{NAc,3S,6S}$	619.51
0	1	1	0	0	C	I_{2S}-H_{NS}	497.41
0	1	1	0	1	D	I_{2S}-$H_{NS,6S}$	577.47
0	1	1	1	0	E	I_{2S}-$H_{NS,3S}$	577.47
0	1	1	1	1	F	I_{2S}-$H_{NS,3S,6S}$	657.53
1	0	0	0	0	-0	G-H_{NAc}	379.33
1	0	0	0	1	-1	G-$H_{NAc,6S}$	459.39
1	0	0	1	0	-2	G-$H_{NAc,3S}$	459.39
1	0	0	1	1	-3	G-$H_{NAc,3S,6S}$	539.45
1	0	1	0	0	-4	G-H_{NS}	417.35
1	0	1	0	1	-5	G-$H_{NS,6S}$	497.41
1	0	1	1	0	-6	G-$H_{NS,3S}$	497.41
1	0	1	1	1	-7	G-$H_{NS,3S,6S}$	577.47
1	1	0	0	0	-8	G_{2S}-H_{NAc}	459.39
1	1	0	0	1	-9	G_{2S}-$H_{NAc,6S}$	539.45
1	1	0	1	0	-A	G_{2S}-$H_{NAc,3S}$	539.45
1	1	0	1	1	-B	G_{2S}-$H_{NAc,3S,6S}$	619.51
1	1	1	0	0	-C	G_{2S}-H_{NS}	497.41
1	1	1	0	1	-D	G_{2S}-$H_{NS,6S}$	577.47

I/G	2X	NX	3X	6X	ALPH CODE	DISACC	MASS (ΔU)
1	1	1	1	0	-E	G_{2S}-$H_{NS,3S}$	577.47
1	1	1	1	1	-F	G_{2S}-$H_{NS,3S,6S}$	657.53

TABLE 5

It has been observed that disaccharide units in some HLGAG sequences are neither N-sulfated nor N-acetylated. Such disaccharide units may be represented using the chemical unit ID 204a in any of a number of ways.

If the properties of a chemical unit are represented by bit fields, disaccharide units that contain a free amine in the N position may be represented by, for example, adding an additional bit field. For example, referring to FIG. 2D, an additional field NY may be used in the chemical unit ID 204a. For example, an NY field having a value of zero may correspond to a free amine, and an NY field having a value of one may correspond to N-acetylation, or vice versa. Further, a value of one in the NX field 212e may correspond to N-sulfation.

Optionally, disaccharide units that contain a free amine in the N position may be represented using a tristate field. For example, the field 212e (NX) in the chemical unit ID 204a may be a tristate field having three permissible values. For example, a value of zero may correspond to a free amine, a value of one may correspond to N-acetylation, and a value of two could correspond to N-sulfation. Similarly, the values of any of the fields 212a-e may be represented using a number system with a base higher than two. For example, if the value of the field 212e (NX) is represented by a single-digit number having a base of three, then the field 212e may store three permissible values.

Referring to Fig. 1, user may perform a query on the polymer database 102 to search for particular information. For example, a user may search the polymer database 102 for specified polymers, specified chemical units, or polymers or chemical units having specified properties. A user may provide to a query user interface 108 user input 106 indicating properties for which to search. The user input 106 may, for example, indicate one or more chemical units, a polymer of chemical units or one or more properties to search for using, for example, a standard character-based notation. The query user interface 108 may, for example, provide a graphical user interface (GUI) which allows the user to select from a list of properties using an input device such as a keyboard or a mouse.

The query user interface 108 may generate a search query 110 based on the user input 106. A search engine 112 may receive the search query 110 and generate a mask 114 based on the search query. Example formats of the mask 114, and example techniques to determine whether properties specified by the mask 114 match properties of polymers in the polymer database 102 are described in more detail below in connection to Fig. 3.

The search engine 112 may determine whether properties specified by the mask 114 match properties of polymers stored in the polymer database 102. Subsequently, the search engine 112 may generate search results 116 based on the search indicating whether the polymer database 102 includes polymers having the properties specified by the mask 114. The search results 116 also may indicate polymers in the polymer database 102 that have the properties specified by the mask 114. For example, if the user input 106 specified properties of a chemical unit, the search results 116 may indicate which polymers in the polymer database 102 include the specified chemical unit. Alternatively, if the user input 106 specified particular chemical unit properties, the search results 116 may indicate polymers in the polymer database 102 that include chemical units having the specified chemical unit properties. Similarly, if the user input 106 specified particular polymer properties, the search results 116 may indicate which polymers in the polymer database 102 have the specified polymer properties.

Fig. 3 is a flowchart illustrating an example of a process 300 that may be used by the search engine 112 to generate the search results 116. In act 302, the search engine 112 may receive a search query 110 from the query user interface 108. Next, in act 304, the search engine 112 may generate a mask 114 generated based on the search query 110. In a following act 306, the search engine 112 may perform a binary operation on one or more of the records 104a-*n* in the polymer database 102 by applying the mask 114. Next, in act 308, the search engine 112 may generate the search results 116 based on the results of the binary operation performed in step 306.

The process 300 will now be described in more detail with respect to an embodiment in which the fields 206a-*m* of the chemical unit 204a are binary fields. In act 302, the received search query 110 may indicate to search the polymer database 102 for a particular chemical unit, e.g. the chemical unit I_{2S}-H_{NS}. If , for example, the coding scheme shown in Table 1 is used to encode chemical units in the polymer database, the chemical unit I_{2S}-H_{NS} may be represented by a binary value of 01001. To generate the mask 114 for this chemical unit (step 304), the search engine 112 may use the binary value of the chemical unit, i.e., 01001, as the value of the mask 114. As a result, the values of the bits of the mask 114 may

specify the properties of the chemical unit I_{2S}-H_{NS}. For example, the value of zero in the leftmost bit position may indicate Iduronic, and the value of one in the next bit position may indicate that the 2X position is sulfated.

The search engine 112 may use this mask 114 to determine whether polymers in the polymer database 102 contain the chemical unit I_{2S}-H_{NS}. To make this determination, the search engine 112 may perform a binary operation on the data units 104a-*n* of the polymer database 102 using the mask 114 (step 306). For example, the search engine 112 may perform a logical AND operation on each chemical unit of each of the polymers in the polymer database 102 using the mask 114. If the result of the logical AND operation on a particular chemical unit is equal to the value of the mask 114, then the chemical unit may satisfy the search query 110, and, in act 308, the search engine 112 may indicate a successful match in the search results 116. The search engine 112 may generate additional information in the search results 116, such as the polymer identifier of the polymer containing the matching chemical unit.

In response to receiving the search query in act 302, in act 304, the search engine 112 also may generate the mask 114 that indicates one or more properties of a particular polymer or chemical unit. To generate the mask 114 for such a search query, the search engine 112 may set each bit position in the mask according to a property specified by the search query to the value specified by the search query. Consider, for example, search query 110 that indicates a search for all chemical units in which both the 2X position and the 6X position are sulfated. To generate a mask corresponding to this search query, the search engine 112 may set the bit positions of the mask corresponding to the 2X and 6X positions to a value corresponding to being sulfated. Using the coding scheme shown above in Table 1, for example, in which the 2X and 6X positions have bit positions of 3 and 2 (counting from the rightmost position beginning at bit position zero), respectively, the mask corresponding to this search query is 01100. The two bits of this mask that have a value of one correspond to the bit positions in Table 1 corresponding to the 2X and 6X positions.

To determine whether the one or more properties of a particular chemical unit in the polymer database 102 match the one or more properties specified by the mask 114, the search engine 112 may perform a logical AND operation on the chemical unit identifier of the chemical unit in the polymer database 102 using the mask 114. To generate search results for this chemical unit (i.e., act 308), the search engine 112 may compare the result of the logical AND operation to the mask 114. If the values of the bit positions of the logical AND operation corresponding to the properties specified by the search query are equal to

the values of the same bit positions of the mask 114, then the chemical unit has the properties specified by the search query 110, and the search engine 112 indicates a successful match in the search results 116.

For example, consider the search query 110 described above, which indicates a search for all chemical units in which both the 2X position and the 6X position are sulfated. Using the coding scheme of Table 1, the bit positions corresponding to the 2X and 6X positions are bit positions 3 and 2. Therefore, after performing a logical AND operation on the chemical unit identifier of a chemical unit using the mask 114, the search engine 112 compares bit positions 3 and 2 of the result of the logical AND operation to bit positions 3 and 2 of the mask. If the values in both bit positions are equal, then the chemical unit has the properties specified by the mask 114.

The techniques described above for generating the mask 114 and searching with a mask 114 also may be used to perform searches with respect to sequences of chemical units or entire polymers. For example, if the search query 110 indicates a sequence of chemical units, the search engine 112 may fill the mask 114 with a sequence of bits corresponding to the concatenation of the binary encodings of the specified sequence of chemical units. The search engine 112 may then perform a binary AND operation on the polymer identifiers in the polymer database 102 using the mask 114, and generate the search results 116 as described above.

The techniques described above for generating the mask 114 and searching with the mask 114 are provided merely as an example. Other techniques for generating and searching with the mask 114 may also be used. The search engine 112 also may use more than one mask for each search query 110, and the search engine 112 may perform multiple binary operations in parallel in order to improve computational efficiency. In addition, binary operations other than a logical AND may be used to determine whether properties of the polymers in the polymer database 102 match the properties specified by the mask 114. Other binary operations include, for example, logical OR and logical XOR (exclusive or). Such binary operations may be used alone or in combination with each other.

Using the techniques described above, the polymer database 102 may be searched quickly for particular chemical units. One advantage of the process 300, if used in conjunction with a chemical unit coding scheme that encodes properties of chemical units using binary values is that a chemical unit identifier (e.g., the chemical unit identifier 204a) may be compared to a search query (in the form of a mask) using a single binary operation (e.g., a binary AND operation). As described above, conventional notation systems that use

character-based notation systems to encode sequences of chemical units (e.g., systems which encode DNA sequences as sequences of characters) typically search for a sub-sequence of chemical units (represented by a first sequence of characters) within a super-sequence of chemical units (represented by a second sequence of characters) and use character-based comparison. Such a comparison typically is slow because it sequentially compares each character in a first sequence of characters (corresponding to the sub-sequence) to characters in a second sequence until a match is found. Consequently, the speed of the search is related to the length of the sub-sequence--i.e., the longer the sub-sequence, the slower the search.

In contrast, the speed of the techniques described above for searching binary operations may be constant in relation to the length of a sub-sequence that is the basis for the search query. Because the search engine 112 can search for a query sequence of chemical units using a single binary operation (e.g., a logical AND operation) regardless of the length of the query sequence, searches may be performed more quickly than conventional character-based methods whose speed is related to the length of the query sequence. Further, the binary operations used by the search engine 112 may be performed more quickly because conventional computer processors are designed to perform binary operations on binary data.

A further advantage of the techniques described above for searching using binary operations is that encoding one or more properties of a polymer into the notational representation of the polymer enables the search engine 112 to quickly and directly search the polymer database 102 for particular properties of polymers. Because the properties of a polymer are encoded into the polymer's notational representation, the search engine 112 may determine whether the polymer has a specified property by determining whether the specified property is encoded in the polymer's notational representation. For example, as described above, the search engine 112 may determine whether the polymer has the specified property by performing a logical AND operation on the polymer's notational representation using the mask 114. This operation may be performed quickly by conventional computer processors and may be performed using only the polymer's notational representation and the mask, without reference to additional information about the properties of the polymer.

Some aspects of the techniques described herein for representing properties using binary notation may be useful for generating, searching and manipulating information about polysaccharides. Accordingly, complete building block of a polymer may be assigned a

unique numeric identifier, which may be used to classify the complete building block. For example, each numeric identifier may represent a complete building block of a polysaccharide, including the exact chemical structure as defined by the basic building block of a polysaccharide and all of its substituents, charges etc. A basic building block refers to a basic ring structure such as iduronic acid or glucuronic acid but does not include substituents, charges etc. Such building block information may be generated and processed in a same or similar manner as described above with respect to "properties" of polymers.

A computer system that may implement the system 100 of FIG. 1 as a computer program typically may include a main unit connected to both an output device which displays information to a user and an input device which receives input from a user. The main unit generally includes a processor connected to a memory system via an interconnection mechanism. The input device and output device also may be connected to the processor and memory system via the interconnection mechanism.

One or more output devices may be connected to the computer system. Example output devices include a cathode ray tube (CRT) display, liquid crystal displays (LCD), printers, communication devices such as a modem, and audio output. One or more input devices also may be connected to the computer system. Example input devices include a keyboard, keypad, track ball, mouse, pen and tablet, communication device, and data input devices such as sensors. The subject matter disclosed herein is not limited to the particular input or output devices used in combination with the computer system or to those described herein.

The computer system may be a general purpose computer system which is programmable using a computer programming language, such as C++, Java, or other language, such as a scripting language or assembly language. The computer system also may include specially-programmed, special purpose hardware such as, for example, an Application-Specific Integrated Circuit (ASIC). In a general purpose computer system, the processor typically is a commercially-available processor, of which the series x86, Celeron, and Pentium processors, available from Intel, and similar devices from AMD and Cyrix, the 680X0 series microprocessors available from Motorola, the PowerPC microprocessor from IBM and the Alpha-series processors from Digital Equipment Corporation, are examples. Many other processors are available. Such a microprocessor executes a program called an operating system, of which Windows NT, Linux, UNIX, DOS, VMS and OS8 are examples, which controls the execution of other computer programs and provides scheduling, debugging, input/output control, accounting, compilation, storage assignment, data

management and memory management, and communication control and related services. The processor and operating system define a computer platform for which application programs in high-level programming languages may be written.

A memory system typically includes a computer readable and writeable nonvolatile recording medium, of which a magnetic disk, a flash memory and tape are examples. The disk may be removable, such as a "floppy disk," or permanent, known as a hard drive. A disk has a number of tracks in which signals are stored, typically in binary form, i.e., a form interpreted as a sequence of one and zeros. Such signals may define an application program to be executed by the microprocessor, or information stored on the disk to be processed by the application program. Typically, in operation, the processor causes data to be read from the nonvolatile recording medium into an integrated circuit memory element, which is typically a volatile, random access memory such as a dynamic random access memory (DRAM) or static memory (SRAM). The integrated circuit memory element typically allows for faster access to the information by the processor than does the disk. The processor generally manipulates the data within the integrated circuit memory and then copies the data to the disk after processing is completed. A variety of mechanisms are known for managing data movement between the disk and the integrated circuit memory element, and the subject matter disclosed herein is not limited to such mechanisms. Further, the subject matter disclosed herein is not limited to a particular memory system.

The subject matter disclosed herein is not limited to a particular computer platform, particular processor, or particular high-level programming language. Additionally, the computer system may be a multiprocessor computer system or may include multiple computers connected over a computer network. It should be understood that each module (e.g. 110, 120) in FIG. 1 may be separate modules of a computer program, or may be separate computer programs. Such modules may be operable on separate computers. Data (e.g., 104, 106, 110, 114 and 116) may be stored in a memory system or transmitted between computer systems. The subject matter disclosed herein is not limited to any particular implementation using software or hardware or firmware, or any combination thereof. The various elements of the system, either individually or in combination, may be implemented as a computer program product tangibly embodied in a machine-readable storage device for execution by a computer processor. Various steps of the process may be performed by a computer processor executing a program tangibly embodied on a computer-readable medium to perform functions by operating on input and generating output. Computer programming languages suitable for implementing such a system include

procedural programming languages, object-oriented programming languages, and combinations of the two.

Having now described a few embodiments, it should be apparent to those skilled in the art that the foregoing is merely illustrative and not limiting, having been presented by way of example only. Numerous modifications and other embodiments are within the scope of one of ordinary skill in the art and are contemplated as falling within the scope of the invention.

What is claimed is:

Claims

1. A data structure, tangibly embodied in a computer-readable medium, representing a polymer of chemical units, the data structure comprising:

an identifier including one or more fields, each field for storing a value corresponding to one or more properties of the polymer,

wherein at least one field stores a non-character-based value.

2. The data structure of claim 1, wherein each of the fields is capable of storing a binary value.

3. The data structure of claim 1, wherein the identifier is representable as a single-digit hexadecimal number.

4. The data structure of claim 1, wherein the identifier is representable as a decimal value.

5 The data structure of claim 4, wherein the decimal value may be reduced to a plurality of prime divisors, wherein each prime divisor represents a building block of the polymer.

6. The data structure of claim 1, wherein the polymer of chemical units comprises a polysaccharide and wherein each of the chemical units is a saccharide.

7. The data structure of claim 1, wherein the polymer of chemical units comprises a nucleic acid and wherein each of the chemical units is a nucleotide.

8. The data structure of claim 1, wherein the polymer of chemical units comprises a polypeptide and wherein each of the chemical units is an amino acid.

9. The data structure of claim 1, wherein the one or more properties comprise one or more chemical unit properties, each chemical unit property being a property of one of the chemical units of the polymer.

10. The data structure of claim 9, wherein the one or more properties comprise one or more charges, each charge being a charge of one of the chemical units of the polymer.

11. The data structure of claim 9, wherein the one or more properties comprise one or more chemical unit identities, each chemical unit identity being an identity of a chemical unit of the polymer.

12. The data structure of claim 9, wherein the one or more properties comprise one or more confirmations, each confirmation being a confirmation of a chemical unit of the polymer.

13. The data structure of claim 9, wherein the one or more properties comprise one or more substituent identities, each substituent identity being an identity of a substituent of a chemical unit of the polymer.

14. The data structure of claim 1, wherein the one or more properties comprise one or more properties of the polymer.

15. The data structure of claim 14, wherein the one or more properties comprise a total charge of the polymer.

16. The data structure of claim 14, wherein the one or more properties comprise a total number of sulfates of the polymer.

17. The data structure of claim 14, wherein the one or more properties comprise a dye-binding of the polymer.

18. The data structure of claim 14, wherein the one or more properties comprise one or more properties of a polysaccharide.

19 The data structure of claim 18, wherein the one or more properties of a polysaccharide include one or more compositional ratios of substituents.

20. The data structure of claim 18, wherein the one or more properties of a polysaccharide include one or more compositional ratios of iduronic versus glucuronic.

21. The data structure of claim 18, wherein the one or more properties of a polysaccharide include enzymatic sensitivity.

22. The data structure of claim 14, wherein the one or more properties comprise a mass of the polymer.

23. The data structure of claim 14, wherein the one or more properties comprise degree of sulfation.

24. The data structure of claim 14, wherein the one or more properties comprise charge.

25. The data structure of claim 14, wherein the one or more properties comprise chirality.

26. The data structure of claim 1, wherein the identifier comprises a numerical identifier.

27. A computer-implemented method for generating a data structure, tangibly embodied in a computer-readable medium, representing a polymer of chemical units, the method comprising an act of:

generating an identifier including one or more fields for storing values, each value corresponding to one or more properties of the polymer,

wherein at least one field stores a non-character-based value.

28. A computer-implemented method for determining whether properties of a query sequence of chemical units match properties of a polymer of chemical units, the query sequence being represented by a first data structure, tangibly embodied in a computer-readable medium, including an identifier that includes one or more fields, each field storing a value corresponding to one or more properties of the query sequence, the polymer being represented by a second data structure, tangibly embodied in a computer-readable medium, including an identifier that includes one or more fields, each field for storing a value corresponding to one or more properties of the polymer, the method comprising acts of:

(A) generating at least one mask based on the values stored in the one or more fields of the first data structure;

(B) performing at least one binary operation on the values stored in the one or more fields of the second data structure using the at least one mask to generate at least one result; and

(C) determining whether the one or more properties of the query sequence match the one or more properties of the polymer based on the at least one result.

29. The method of claim 28, wherein each of the one or more fields of the first and second data structures is a bit field.

30. The method of claim 28, wherein the act (A) comprises an act of:

(A)(1) generating the at least one mask as a sequence of bits that is equivalent to the values stored in the fields of the first data structure.

31. The method of claim 28, wherein the act (A) comprises an act of:

(A)(1) generating the at least one mask as a sequential repetition of the values stored in the fields of the first data structure.

32. The method of claim 28, wherein the at least one mask comprises a plurality of masks and wherein the act (B) comprises acts of:

(B)(1) performing a logical AND operation on the values stored in the fields of the second data structure using each of the plurality of masks to generate a plurality of intermediate results; and

(B)(2) combining the plurality of intermediate results using at least one logical OR operation to generate the at least one result.

33. The method of claim 28, wherein the act (C) comprises an act of:

(C)(1) determining that the one or more properties of the query sequence match the one or more properties of the polymer when the at least one result has a non-zero value.

34. The method of claim 28, wherein the at least one binary operation comprises at least one logical AND operation.

35. A database, tangibly embodied in a computer-readable medium, for storing information descriptive of one or more polymers, the database comprising:

one or more data units corresponding to the one or more polymers, each of the data units including an identifier that includes one or more fields, each field for storing a value corresponding to one or more properties of the polymer.

36. A method for determining whether complete building blocks of a query sequence of chemical units match complete building blocks of a polysaccharide, the query sequence being represented by a first data structure, tangibly embodied in a computer-readable medium, including an identifier that includes one or more fields, each field for storing a value corresponding to a complete building block of the query sequence, the polysaccharide being represented by a second data structure, tangibly embodied in a computer-readable medium, including an identifier that includes one or more fields, each field for storing a value corresponding to a complete building block of the polysaccharide, the method comprising acts of:

(A) generating at least one mask based on the values stored in the one or more fields of the first data structure;

(B) performing at least one binary operation on the values stored in the one or more fields of the second data structure using the at least one mask to generate at least one result; and

(C) determining whether the complete building blocks of the query sequence match the complete building blocks of the polysaccharide based on the at least one result.

37. The method of claim 36, wherein each of the one or more fields of the first and second data structures is a bit field.

38. A data structure, tangibly embodied in a computer-readable medium, representing a polysaccharide, the data structure comprising:

an identifier including one or more fields, each field for storing a value corresponding to a complete building block of the polysaccharide.

39. The data structure of claim 38, wherein each of the one or more fields are capable of storing a binary value.

40. The data structure of claim 38, wherein the identifier is representable as a single-digit hexadecimal number.

41. The data structure of claim 38, wherein the identifier is representable as a decimal value.

42 The data structure of claim 41, wherein the decimal value can be reduced to a plurality of prime divisors, wherein each prime divisor represents a building block of the polysaccharide.

43. A data structure, tangibly embodied in a computer-readable medium, representing a chemical unit of a polymer, the data structure comprising:

an identifier including one or more fields, each field for storing a value corresponding to one or more properties of the chemical unit,

wherein at least one field stores a non-character-based value.

44. The data structure of claim 43, wherein the one or more properties include a charge of the chemical unit.

45. The data structure of claim 43, wherein the one or more properties include an identity of the chemical unit.

46. The data structure of claim 43, wherein the one or more properties include a confirmation of the chemical unit.

47. The data structure of claim 43, wherein the one or more properties include an identity of a substituent of the chemical unit.

48 The data structure of claim 43, wherein each of the fields is capable of storing a binary value.

49. The data structure of claim 43, wherein the identifier is representable as a single-digit hexadecimal number.

50. The data structure of claim 43, wherein the identifier is representable as a decimal value.

51 The data structure of claim 50, wherein the decimal value is a primary number.

52. The data structure of claim 51, wherein the polymer is a polysaccharide, and the primary number identifies the chemical unit as a building block of the polysaccharide.

53. The data structure of claim 43, wherein the polymer is a polysaccharide.

Abstract

A data structure, tangibly embodied in a computer-readable medium, representing a polymer of chemical units is disclosed. The data structure includes an identifier including a plurality of fields for storing values corresponding to properties of the polymer. In one embodiment, the fields are capable of storing binary values. The polymer may, for example, be a polysaccharide and the chemical units may be saccharides. Also disclosed is a computer-implemented method for determining whether properties of a query sequence of chemical units match properties of a polymer of chemical units. The query sequence is represented by a first data structure, tangibly embodied in a computer-readable medium, including an identifier including a plurality of bit fields for storing values corresponding to properties of the query sequence. The polymer is represented by a second data structure, tangibly embodied in a computer-readable medium, including an identifier including a plurality of bit fields for storing values corresponding to properties of the polymer. The method includes steps of generating at least one mask based on the values stored in the bit fields of the first data structure; performing at least one binary operation on the values stored in the bit fields of the second data structure using the at least one mask to generate at least one result; and determining whether the properties of the query sequence match the properties of the polymer based on the at least one result. The invention also involves a notational system referred to as Property Encoded Nomenclature.

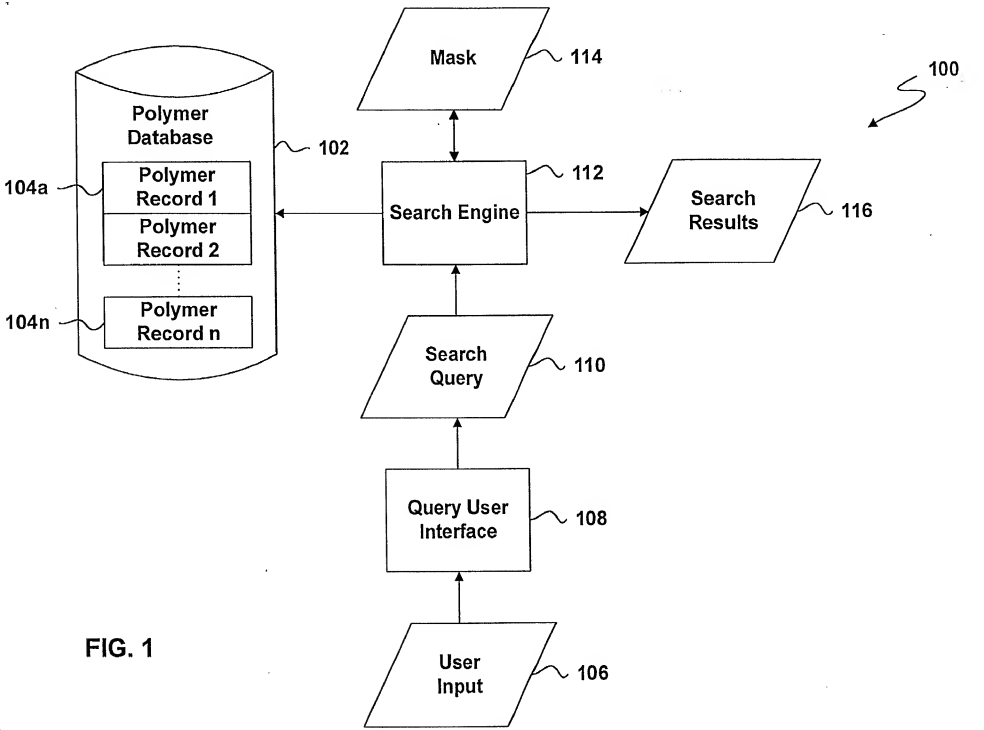
Mask
114
100
Polymer
Database
102
104a
Polymer
Record 1
Polymer
Record 2
104n
Polymer
Record n
Search Engine
112
Search
Results
116
Search
Query
110
Query User
Interface
108
User
Input
106

FIG. 1

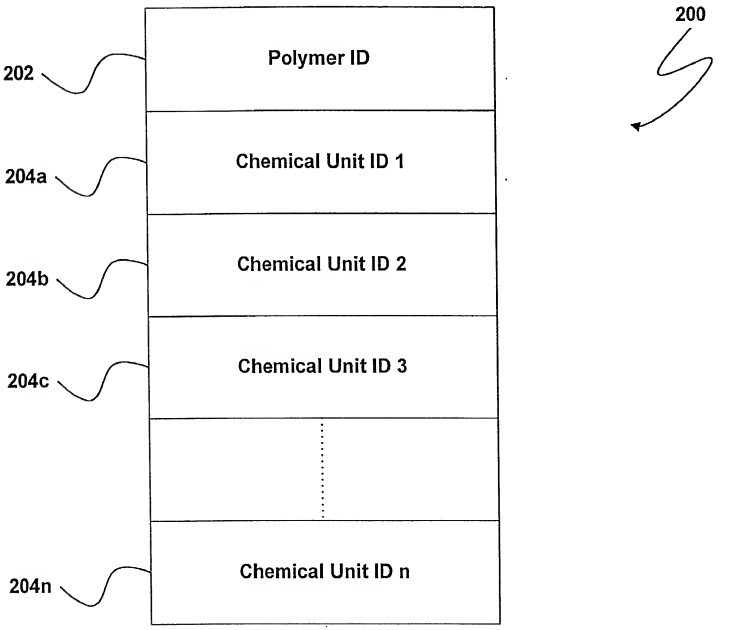
200
202
Polymer ID
204a
Chemical Unit ID 1
204b
Chemical Unit ID 2
204c
Chemical Unit ID 3
204n
Chemical Unit ID n

FIG. 2A

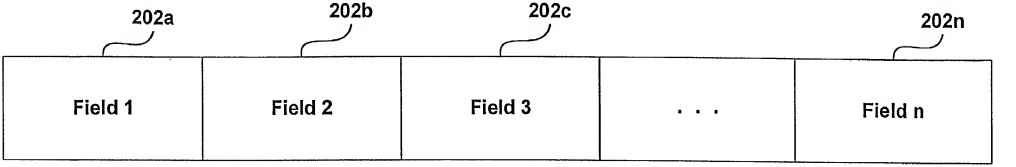
202a
202b
202c
202n
Field 1
Field 2
Field 3
. . .
Field n
202

FIG. 2B

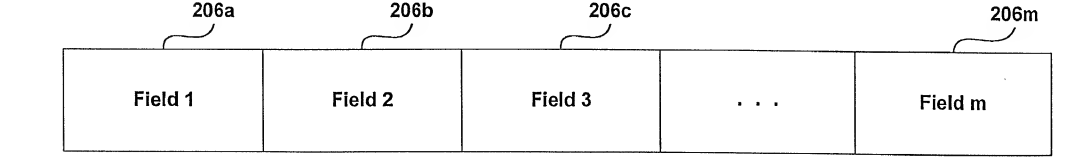
206a
206b
206c
206m
Field 1
Field 2
Field 3
. . .
Field m
204a

FIG. 2C

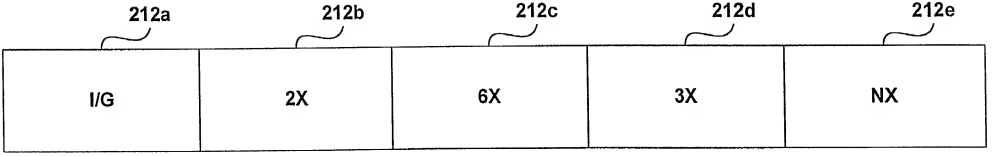

FIG. 2D

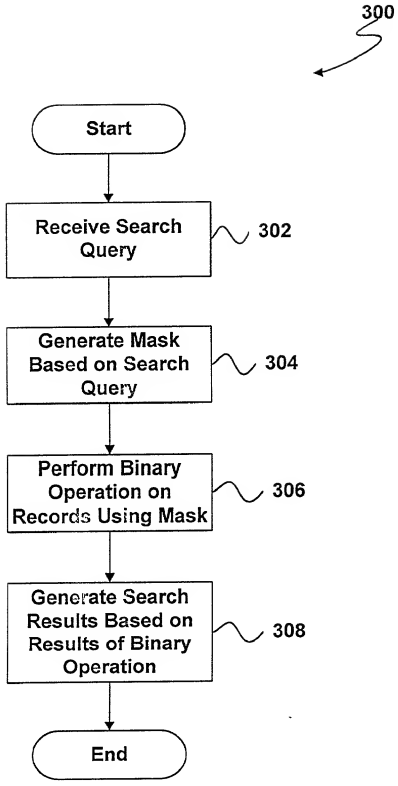
300
Start
Receive Search Query
302
Generate Mask Based on Search Query
304
Perform Binary Operation on Records Using Mask
306
Generate Search Results Based on Results of Binary Operation
308
End

FIG. 3